A QUICK METHOD FOR DETERMINING THE INDEX OF CORRELATION.

By GUY MONTROSE WHIPPLE, Ph. D.
Assistant Professor of Education, Cornell University.

The desirability of substituting an accurate numerical index for mere verbal expressions of correlation has been very clearly set forth by Galton, Pearson, Yule, Spearman, Wissler, and other writers.[1] But the most accurate formula, the 'product-moments' formula of Pearson, is attended with arduous labor. We may greatly abridge the numerical work by the use of an adding machine[2] and of appropriate tables, such as Barlow's *Table of Squares*, and Krelle's *Multiplication Tables*. Further, the computation of σ, the standard deviation, may often be reduced by considering it as equal to *m. v.*, or the average deviation, times the constant, 1.2533. Yet, even so, the task is considerable, so that, particularly if one has to determine a number of correlations, it is desirable to use a shorter method for preliminary exploration.

A number of shorter correlation methods have been described.[3] It is the purpose of the present article to describe a simplification of one of these methods that the writer has found very expeditious and serviceable for the determination of an approximate numerical correlation. This method is based upon the use of what is known as Sheppard's formula, which may itself be regarded as a simplification of one of Pearson's auxiliary methods.

For the application of this, as of most formulas, the data of

[1]See, for instance, F. Galton, *Natural Inheritance;* C. Spearman, The Proof and Measurement of Association Between Two Things, this *Journal*, XV, 1904, 72; C. Wissler, The Correlation of Mental and Physical Tests, *Psych. Rev. Mon. Supp.* No. 16, 1901. The important contributions of K. Pearson, and R. Yule, will be found in the *Proc. Royal Soc.* of London, in the *Phil. Transactions* of the same body, the *Jour. Royal Stat. Soc.*, and, in their more recent applications to biological problems, in the several volumes of the *Biometrika*, 1901 ff.

[2]An inexpensive, but very serviceable device, known as the Gem Adder, is now put on the market by the Automatic Adding Machine Company, Broome St., New York City, at a price of fifteen dollars, and is well worth purchase by any one who contemplates correlation work.

[3]See, for example, the articles by Spearman and Wissler.

each of the two series to be compared must first be distributed in an orderly array. Suppose, to take a concrete instance, we wish to ascertain the correlation between the accuracy with which 50 boys cancel *e* from a printed slip and the accuracy with which the same 50 boys cancel *q*, *r*, *s* and *t* from a similar slip. The results of each test are first arranged in order, the least accurate boy first and the most accurate last. We can then either determine the average, in which case all the boys that rank below the average are minus and all that rank above are plus, or we can simply take the median value and consider the first 25 boys in each array as minus, and the second 25 as plus, cases. By rapid comparison the following values are next determined:

a = no. cases that are plus in the 1st and plus in the 2d series.
b = " " " " plus " " " " minus " " " "
c = " " " " minus " " " " plus " " " "
d = " " " " minus " " " " minus " " " "

The index of correlation may now be obtained by reference to one of Pearson's simpler formulas:

$$r = \sin \frac{\pi}{2} \frac{\sqrt{ad} - \sqrt{bc}}{\sqrt{ad} + \sqrt{bc}}$$

Now this formula may be brought into a more convenient form if we replace the sine by the cosine of its complement.

$$r = \cos \left[\frac{\pi}{2} - \frac{\pi}{2} \frac{\sqrt{ad} - \sqrt{bc}}{\sqrt{ad} + \sqrt{bc}} \right]$$

when we can reduce to

$$r = \cos \frac{\sqrt{bc}}{\sqrt{ad} + \sqrt{bc}} \pi.$$

If, now, we further simplify by substituting for the square root of the product of the *b* and *c* cases the percentage of cases with unlike signs (U), and for the square root of the product of the *a* and *d* cases the percentage of cases with like signs (L),[1] we obtain Sheppard's formula:

$$r = \cos \frac{U}{L + U} \pi$$

The results of this formula do not differ appreciably from the foregoing as the value of the fraction is virtually identical.

[1] That is, virtually, substituting the arithmetical for the geometrical mean.

Now, since L + U must always equal 100, and since $\pi = 180°$, this formula may be written for greater convenience,

$$r = \cos U\ 1.8°$$

Finally, since the values of U must range from 50 to 0 for positive, and from 50 to 100 for inverse correlations, it now becomes possible to prepare a simple table from which the values of r for any integer value of U may be read directly, and I have here introduced this Table in the hope that it may prove of interest and assistance.

Correlation Table.

for the formula $r = \cos U\ 1.8°$

If U is greater than 50, first substract it from 100, then prefix the minus sign to the correlation indicated.

U	r	U	r	U	r	U	r	U	r
0	1.000	10	.951	20	.809	30	.587	40	.309
1	.999	11	.941	21	.790	31	.562	41	.279
2	.998	12	.929	22	.770	32	.536	42	.248
3	.995	13	.917	23	.750	33	.509	43	.218
4	.992	14	.904	24	.728	34	.482	44	.187
5	.987	15	.891	25	.707	35	.454	45	.156
6	.982	16	.876	26	.684	36	.426	46	.125
7	.976	17	.860	27	.661	37	.397	47	.094
8	.968	18	.844	28	.637	38	.368	48	.062
9	.960	19	.827	29	.613	39	.338	49	.031

It will be seen, then, that the discovery of the approximate numerical value of the correlation between two series of data is reduced to four simple steps, (1) distribution of the data into two arrays sectioned at the median, (2) counting the cases with unlike signs and (3) dividing this number by the total number of cases, (4) reference to the Table.

The probable error may be calculated from the formula:[1]

$$p.\ e. = \sin\left[0.1686\ \pi\ (1 - r^2)\sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}\right]$$

To illustrate the employment of these methods, in the example cited the following values were obtained: $a = 18$, $b = 11$, $c = 8$, $d = 13$. Hence $U = 38$. By the use of either short formula, $r = +.37$ with p. e. $= .26$. By the use of Pearson's product-moments method we obtain for the same arrays, $r = .47$, with $p.\ e. = .06$, but, by actual timing, after the distributions had been made the first method occupied eight minutes

[1] The value of $(1-r^2)$ for all values of r may be obtained directly from a Table published by Yule, *Jour. R. S. Soc.*, X., 1897, 852-3.

and the second two hours and fifteen minutes, even with the aid of the adding machine and the Tables previously mentioned.

The short method cannot, of course, be recommended for the final determination of important correlations, because the probable error is large, particularly with relatively few cases and a low value of r, but it is very serviceable for the preliminary examination of such data, and may give results of value when the scale divisions are fairly fine, the data symmetrically distributed, the number of cases not too small, and the correlation large, say above 0.50.